W12–W15 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm221

eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications

Mounir Errami[1], Jonathan D. Wren[2], Justin M. Hicks[1] and Harold R. Garner[1,*]

[1]McDermott Center for Human Growth and Development and the Department for Translational Research, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9185 and [2]Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation; 825 N.E. 13th Street, Oklahoma City, Oklahoma 73104-5005, USA

Received January 29, 2007; Revised March 15, 2007; Accepted March 28, 2007

ABSTRACT

Authors, editors and reviewers alike use the biomedical literature to identify appropriate journals in which to publish, potential reviewers for papers or grants, and collaborators (or competitors) with similar interests. Traditionally, this process has either relied upon personal expertise and knowledge or upon a somewhat unsystematic and laborious process of manually searching through the literature for trends. To help with these tasks, we report three utilities that parse and summarize the results of an abstract similarity search to find appropriate journals for publication, authors with expertise in a given field, and documents similar to a submitted query. The utilities are based upon a program, eTBLAST, designed to identify similar documents within literature databases such as (but not limited to) MEDLINE. These services are freely accessible through the Internet at http://invention.swmed.edu/etblast/etblast.shtml, where users can upload a file or paste text such as an abstract into the browser interface.

INTRODUCTION

Searching for pertinent literature is an essential part of every scientist's life. There are many stages in the scientific process in which intimate knowledge of the appropriate literature is critical: (i) familiarization of a new area by a young scientist or a scientist whose research is taking on a new direction, (ii) monitoring the literature as the research progresses to capitalize on recent developments, measure ones competitiveness and avoid duplication of effort (1), (iii) development of reference lists during manuscript or grant application writing and (iv) compiling suggested reviewers when called upon to do so as part of a manuscript submission to a journal. For mature scientists, the reasons for interaction with the literature expand: (i) development of very broad knowledge when writing, for example, a review article, (ii) mastery of new areas in the role of student mentor or examiner and (iii) acquiring focused knowledge when called upon as a manuscript or grant application reviewer. For other scientific professionals, the literature is a resource for identifying colleagues: (i) identification of experts for advisory or steering committees, (ii) selection of reviewers for grants or proposals by government or private agencies, (iii) identification of experts for legal proceedings and testimony, (iv) finding starting points into the literature for novice or lay individuals by librarians and (v) identification of manuscript reviewers by journal editors.

The primary portal for the biomedical literature is PubMed (2,3). This web-based tool searches the Medline database using keywords and Boolean operators. The selection of appropriate keywords by the user requires some knowledge to choose wisely, and this often requires numerous iterations to sample the literature with hopes of finding the most relevant literature. Once the results of a query are presented to the user, the lists can be sorted by date, author or journal. Recent research has focused upon improving the quality and navigation of output (4–8).

There is sufficient information contained within the Medline database to overcome these limitations given a tool with appropriate query entry and result presentation methods. Scientists or professionals either generate in the course of manuscript or grant writing or are presented with concentrated information in the form of an abstract or other document. Given this, the keyword selection and optimization process can be bypassed if natural language free text, such as an abstract, can be submitted directly to a literature search engine. To do this, we have developed eTBLAST, which uses a hybrid scheme to extract and weight keywords contained within the submitted query to identify a subset of literature in Medline, and then performs a sentence alignment to compute a final

*To whom correspondence should be addressed. Tel: 214 648 1661; Fax: 214 648 1445; Email: harold.garner@utsouthwestern.edu

quantitative score as a measure of similarity and, presumably, relevancy. This tool then outputs a list, similar to PubMed, but ranked instead by this similarity score. At this point, scientists can interact with the most relevant Medline literature much as they have done traditionally via date, author or journal sorting methods in PubMed. This similarity-ranked output can be further processed to compile lists and present output views which add value for the specific uses just outlined; identifying the most frequent and prominent authors as experts/reviewers, identifying the most frequent journals as targets for submission and inspection of the publication rate over time as a measure of novelty and topic popularity. It should be noted that eTBLAST and PubMed both find similar abstracts, but by different methods and PubMed's Related Links is limited to only finding similarity among the records currently in Medline, not arbitrary text, as is used by eTBLAST. There also are numerous other Medline keyword-based search tools (CiteXplore, HubMed and GoPubMed, for example) (8–10), including some of which have results post processors with some similar functionality (author and journal finding).

Summarized herein are a set of parsers for the code, eTBLAST (11,12), that can take an abstract or any text as input to identify lists of 'experts', target journals and publication trends.

INPUTTING DATA AND ACCESSING RESULTS

The server requires a text specimen that can be input via copy/paste, or by uploading a text-only file. Additionally an email input option is available to allow users to receive a URL pointing to the results. Results are stored for at least 1 month. The analysis is currently performed on a 20 CPU Linux cluster. The eTBLAST webserver has been up since 2003 and typical searches (of abstracts containing 100–200 words) against Medline, which currently has >16 million records, usually takes from 1 to 3 min and is roughly proportional to the query length. Although Medline is expanding by about 500 000 records per year, eTBLAST performance is continuously being improved through code optimizations and expansion of the number of CPUs in the cluster. There is also a backup 20 CPU Linux cluster which mirrors the primary cluster to guarantee high availability.

eTBLAST [see (9) for a detailed description of methods and performance statistics] returns a list of PubMed IDs (PMID) ordered by statistical similarity to the input text. Briefly, using a two-step process, eTBLAST computes a quantitative score. In step one a weighted keyword set extracted from the query is used to quickly search a database of indexed keywords in Medline, gathering the top 400 most similar records. In step two, a novel sentence alignment algorithm is used to refine the rank order of those similar records and compute a z-score. Each of the utilities presented herein performs a similar set of tasks on these results: (i) results are parsed to extract relevant articles (with similarity z-score $>$ 3), (ii) authors or journals which are overrepresented are calculated and (iii) the results are returned to the user (Figure 1A).

On January 17, 2007 at 17:40 the abstract from (13) was submitted to eTBLAST via the web browser at http://invention.swmed.edu/etblast/index.shtml. Results were returned after 120 s. The query text contained 149 words, of which 58 were 'stop words'. A collage of some of the output web pages is presented in Figure 1, discussed above, to illustrate the output user interface.

Find an expert

Potential reviewers are those who have published frequently in areas highly similar to the query. An author's name may appear on many citations in many different formats, so the last name and first initial are used. An 'Expertise score' is computed for each author that appears in any of the Medline records with an eTBLAST similarity z-score >3, and this Expertise score is used to generate a ranked list which is output to the user. The Expertise score for each author is computed as the sum over all records (1 to N) with a z-score >3:

$$Expertise\ score = \sum w_a{}^* w_p$$

where an arbitrary weight, w_a, is assigned based on the author's position on the author list: the senior author (often last or only author on the list) receives a weight of 3, the lead (first) author 2, and contributing authors 1. 'Corresponding authors,' perhaps a good indicator of expertise, are not explicitly tagged in PubMed, and therefore cannot be used. Each record is also assigned a weight, w_p, which is its similarity score normalized to the query's self-identity similarity score.

To distinguish between 'true' experts and those authors that appear at low frequency within the author lists of the highly similar records, we computed several Expertise score distributions to identify a threshold score. Two sets of queries, each containing 1000 members, were used as input to eTBLAST. The first test set consisted of 1000 Medline records randomly selected from all of Medline. The second test set of 1000 pseudo-random queries, generated with keywords randomly picked from Medline, with the same size distribution and word frequency distribution as Medline, were synthesized using the built-in Perl pseudo-random number generator [as described in (11)]. The top scoring authors (experts) were recorded and the score frequency distributions are presented in Figure 2. From these distributions, we were able to define an Expertise score threshold of 0.9, above which authors can be considered as having the relevant skills (based on their publication history) to be potential experts. This threshold is output on the expert list (Figure 1B). Finally, an expert with no publication in the last 10 years will be flagged as potentially inactive (retirement, change in focus, death, etc...).

Find a journal

The Journal Finder utility parses the user's eTBLAST results in a manner similar to the Expert Finder described

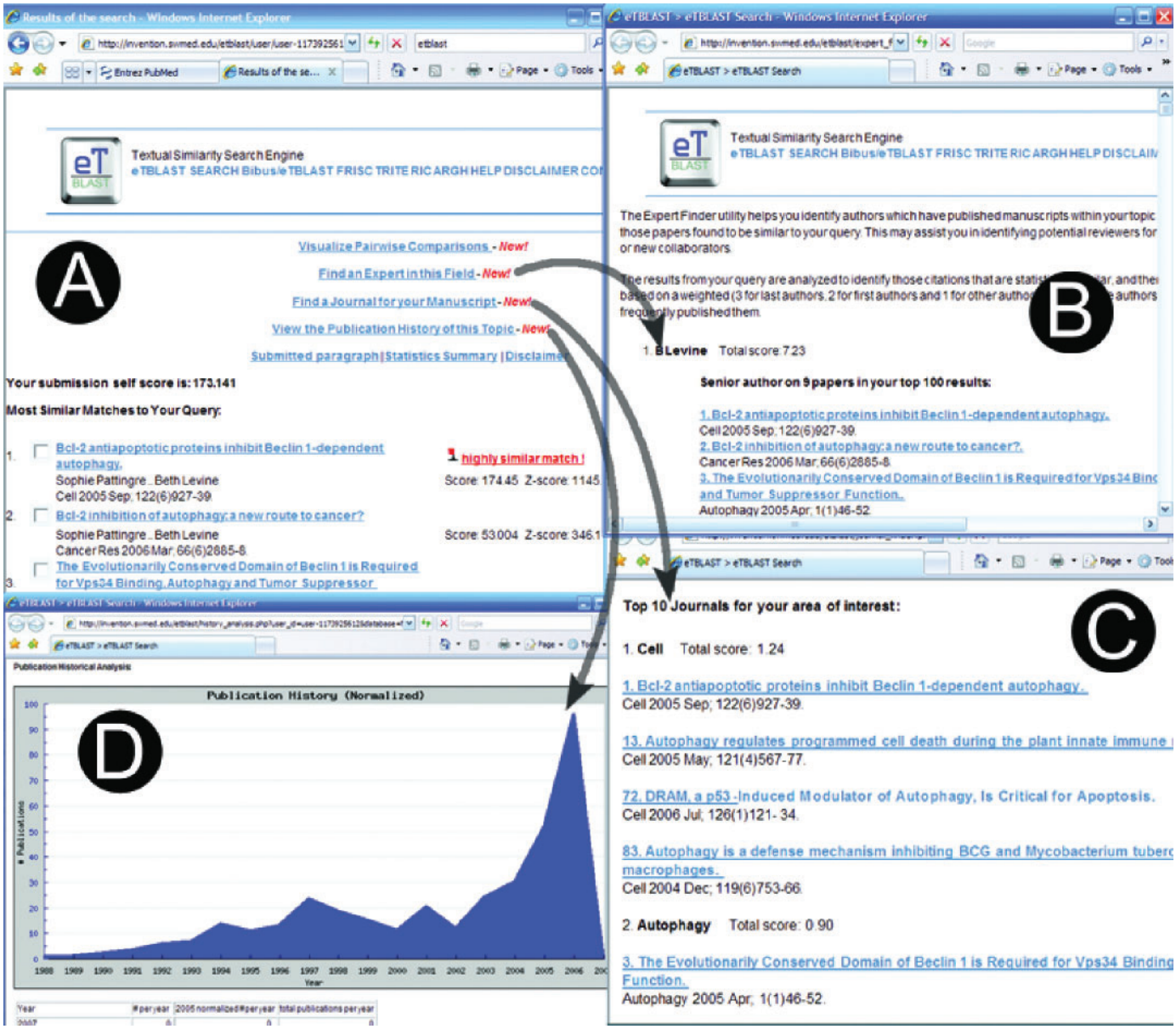

Figure 1. Example output for similarity search (A), experts (B), journals (C) and history of publications (D) obtained with the abstract from PMID16179260 (11) as an illustrative query. The highly similar match flag is raised for the paper from which the abstract was obtained. The top expert is the leading author in the paper from which the abstract was obtained. Finally the first suggested journal is Cell, in which this article was published. These results are biased, since the original paper was not removed from the results. This was done on purpose to illustrate the usage of eTBLAST server.

above (Figure 1B). For those N records with a z-score >3, a Journal score is computed as follows:

$$Journal\ score = \sum w_p$$

where w_p is defined above. The Journal Finder utility lists the highest scoring journals to the web browser ranked by the Journal score and the citations for the publications in that journal. A Journal score threshold, computed similarly to the Expertise score threshold, is also demarked on the output, and is set to 0.1. A benchmark of the Journal Finder utility was conducted using a different set of 4230 abstracts randomly selected from Medline. In 33% of cases, Journal Finder ranked the journal in which the abstract was published within the top 10 suggestions.

Publication history

Authors, reviewers and others can also evaluate the research activity within a given research area as defined in, for example, a manuscript abstract, by the temporal variation of publications found to be similar to the query. For each similar Medline record as found by eTBLAST with a z-score > 3, the year of its publication is parsed from the search results. In this utility the publication year

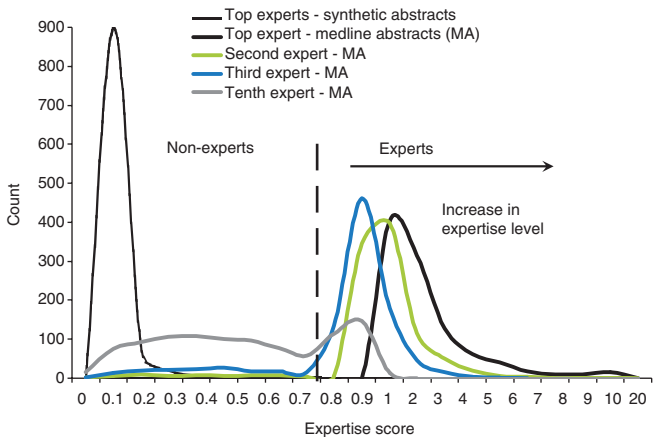

Figure 2. Expertise scores can be used to clearly identify a threshold enabling identification of 'true' experts. A synthetic set of a 1000 abstracts (see text, e.g. non-sensical queries that resemble a typical abstract) was used to determine the score distribution for authors (experts) found in the most similar articles returned by eTBLAST. A second set of 1000 abstracts from randomly selected articles in Medline was used to obtain the score distribution of the first, second, third and tenth authors (experts) on the Find an Expert output list. As the rank of the experts increases, the distribution tends to shift toward the left, with lower scores.

of each record is returned, and a simple count maintained of publications by year. This count is then normalized to the total number of publications for the corresponding year. A tabular output with the raw counts per year and the counts normalized to ratio of the number of publications in Medline in each year divided by the number of publications in 2005 (the basis year) is presented. A graphic is provided for the normalized count over the last 20 years (Figure 1D).

Highly similar publication detection

A score threshold (query self-identity score to similarity score >0.56) has been experimentally defined (unpublished data) and is used to identify and flag any records that are of unusually high similarity to the query as an aid in determining novelty of the topic defined by the query. These either represent abstracts that were taken from Medline for analysis or, if not, serve as a red flag that something very similar to the material being queried has already been published. In our test case, a similarity flag is raised for the paper containing the original abstract (Figure 1A).

CONCLUSIONS

The primary methods in which users interact with the results of Medline searches can be improved and expanded to enable quick and efficient suggestions for optimizing the manuscript writing and publication process, including review. Quantitative similarity scores computed for a text query, such as the abstract for a manuscript submitted to a journal for publication, against the primary biomedical bibliographic database, Medline, can be used to generate a ranked list of similar documents from which summary information about the authors, journals, similar work and dates can be of high utility. Scientists submitting to or editors of the more than 5000 journals represented in Medline can use this free web-based utility to speed the process of selecting or confirming appropriate journal selection, estimate a given articles novelty based on the relative similarity of its abstract and to select potential reviewers (experts), typically requested by journals at manuscript submission time.

Several caveats and potential enhancements to the system should be noted. First, as with any search system, similar articles sharing keywords but belonging to different fields may appear as relevant. Secondly, journal targets or experts are calculated based on frequencies of journals and authors in the eTBLAST results; these suggestions do not account for the publication volume of each journal. Finally, journal impact factors may be indicators of expertise level and are not considered. These enhancements may improve performance and are being evaluated as potential upgrades to the system.

Conflict of interest statement. None declared.

REFERENCES

1. von Elm,E., Poglia,G., Walder,B. and Tramer,M.R. (2004) Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *JAMA*, **291**, 974–980.
2. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **35**, D5–D12.
3. PUBMED interface to Medline, U.S. National Library of Medicine [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed]
4. Goetz,T. and von der Lieth,C. (2005) Pubfinder: a tool for improving retrieval rate of relevant pubmed abstracts. *Nucleic Acids Res.*, **33**, W774–W778.
5. Perez-Iratxeta,C., Perez,A.J., Bork,P. and Andrade,M.A. (2003) Update on xplormed: a web server for exploring scientific literature. *Nucleic Acids Res.*, **31**, 3866–3868.
6. Muin,M. and Fontelo,P. (2006) Technical development of pubmed interact: an improved interface for medline/pubmed searches. *BMC Med. Inform. Decision Making [electronic resource]*, **6**, 36.
7. Ding,J., Hughes,L.M., Berleant,D., Fulmer,A.W. and Wurtele,E.S. (2006) Pubmed assistant: a biologist-friendly interface for enhanced pubmed search. *Bioinformatics*, **22**, 378–380.
8. Eaton,A.D. (2006) Hubmed: a web-based biomedical literature search interface. *Nucleic Acids Res.*, **34**, W745–W747.
9. Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, **33**, W783–W786.
10. CiteXplore helps users to explore literature related to biological research and bio-informatics [http://www.ebi.ac.uk/citexplore/].
11. Lewis,J., Ossowski,S., Hicks,J., Errami,M. and Garner,H.R. (2006) Text similarity: an alternative way to search medline. *Bioinformatics*, **22**, 2298–2304.
12. Pertsemlidis,A. and Garner,H.R. (2004) Text comparison based on dynamic programming. *IEEE Eng. Med. Biol. Mag.: Quart. Mag. Eng. Med. Biol. Soc.*, **23**, 66–71.
13. Pattingre,S., Tassa,A., Qu,X., Garuti,R., Liang,X.H., Mizushima,N., Packer,M., Schneider,M.D. and Levine,B. (2005) Bcl-2 antiapoptotic proteins inhibit beclin 1-dependent autophagy. *Cell*, **122**, 927–939.